信息检索需求描述中的词语区域凸显*)

熊文新[1] 宋 柔[2]
(北京外国语大学中国外语教育研究中心 北京 100089)[1]
(北京语言大学语言信息处理研究所 北京 100083)[2]

摘 要 根据对当前主流信息检索测试 Query 的形式和内容分析,通过正则表达式将 SGML 形式的 Query 表述区分为不同描述域和主题功能块,针对处于不同区域的词语实施不同的加权策略。实验表明,结合主题词语的区域凸显和 TF×IDF 的加权方法比单纯 TF×IDF 方法 MRR 值高出 26.67%。

关键词 信息检索,主题词语,凸显,描述域,功能块

Salience Scheme for Words in Different Parts of Information Request

XIONG Wen-Xin[1] SONG Rou[2]
(National Research Center for Foreign Language Education, Beijing Foreign Studies University, Beijing 100089)[1]
(Center for Language Information Processing, Beijing Language and Culture University, Beijing 100083)[2]

Abstract Based on the analysis of both the form and contents of queries in information retrieval contests, words in queries are partitioned into different descriptive sections and topic chunks by using regular expression matching. Different weighting schemes are applied to words in different descriptive parts. It is shown that a scheme combining the salience of topic words in different sections with TF * IDF outperforms the mere use of TF * IDF. The combination of the methods results in an MRR increase of 26.67%.

Keywords Information retrieval, Topic word, Salient, Descriptive section, Topic chunk

1 引言

文本信息检索是通过分析用自然语言表述信息需求的查询语句(Query),由计算机系统按照某种算法自动从自然语言文本集中检索出符合用户信息需求的特定文本这样一个过程。由于两端都涉及到自然语言,从根本上看这是语言工程问题。

传统向量空间模型采用词袋策略,在过滤停用词后将剩余的所谓内容词语根据 TF * IDF 构造词语权重向量,并以此比较 Query 和目标文本的相似性[1]。这种方法实际上是将不同性质的内容词平等看待,差别只在于词语出现频次的统计。这实质上忽略了不同词语在表述信息内容时所起的作用。我们曾经提出一个自然语言查询问句的处理框架,试图区分用户信息需求表述中的不同性质词语,以期提升信息检索系统的准确率[6]。

用户信息需求的表述多种多样:从 Web 搜索中的一个简单的词语,到提交成篇章的文本。此处,我们限定用户查询表述为当前主流信息检索测试(TREC、NTCIR、CLEF、863IR)中的问题集形式,试图根据用户提交的信息内容词语在测试问题集所处的不同区域,凸显用户的主题信息需求。

2 用户提交问题集 Query 分析

TREC 等文本信息检索测试的资源一般由大规模测试语料和专家人工构造的查询问题集构成。其中,查询问题集又由若干查询语句(Query)组成。Query 是专家模拟日常用户信息需求利用自然语言编写的提问问题,多数采用 SGML 标注,形式上至少包括三个描述域:Title、Description、Narrative。前两者是对 Query 主题短语层面的简短说明,后者是对检索主题成段落的长句描述。典型示例见图 1。

```
<query id="1">
<title>反垃圾邮件措施</title>
<description>防范以及制止垃圾邮件的有关措施</ description >
<narrative>垃圾邮件已经成为令每个人以及政府机构头痛的严重问题。检索防范垃圾邮件的技术措施以及政府相关法律条例,只介绍垃圾邮件及其危害的文章或者有关产品介绍不在检索范围内</ narrative >
</query>
```

图 1 TREC 类 Query 表述

*)国家自然科学基金项目(60272055);国家 863 项目(2001AA114111);教育部科学技术研究重点资助项目(00128)。熊文新 博士后,主研方向:中文信息处理、语言工程;宋 柔 教授,博士生导师,主研方向:智能软件工具、语言信息处理、人工智能。

TREC目前已经成为国际信息检索测试的标杆，已经有很多研究者注意到处于不同描述域的内容词语对确定用户Query主题有重要的指示作用，结合对较复杂的长句描述Narrative域划分不同功能块，选取特定域或特定功能块中的合适词语作为检索项(term)，以提升Query与目标文档的主题内容匹配的相关性处理[2,8]。由于TREC Query中的Narrative域相对简单，它们多数只将其二分为相关条件和不相关条件两部分。而我国863IR测试集在Narrative域还引入检索主题的背景成分，因此对自成段落以长句形式描述的Narrative域有必要做出进一步的区分。

2.1 Query形式切分

Query遵循SGML规范。每个域信息内容采用前后匹配的域描述标记(tag)标示。如图1所示，每个Query有且只有三个域。以FIELD_NAME作为域名标识(field name)，则有

$FIELD\_NAME::=title|description|narrative$

利用正则表达式/〈$FIELD\_NAME$〉(.*)〈\$FIELD_NAME$〉/，提取处于两个配对域名标识之间的信息内容，创建并写入与该域名标记同名的文档。Narrative文档内容相对比较复杂，一般由多个句子构成，需要形成结构化数据。我们通过人工提取总结有关检索范围表述的模式，形成若干区别规则，其形式如下：

(.*)[检索|查询](.*)[等内容|相关文档]，(.*)视为无关。

(.*)在检索范围[之]内。(.*)不在检索范围内。

(.*)与此相关[，|。](.*)与此无关。

相关文件[应|必须][提及|提到|包括](.*)，(.*)则[为|属]非相关文件

……

其中(.*)表示匹配任何字符串；[$x|y$]表示x与y可选出现。

同样，利用正则表达式匹配模式编写脚本文件，将Narrative域中用户希望查询和排除的检索内容分离。对2004年863IR测试问题集的30个Query进行封闭测试，结果如表1所示。

表1 Narrative子文档相关检索需求切分结果

文本总数 (N)	返回结果数 ($Return$)	正确结果数 ($Correct$)	查全率 $P=\frac{Return}{N}$	查准率 $R=\frac{Correct}{Return}$
30	30	28	100%	93.33%

通过这一步骤处理，可区分Narrative域用户真正需要检索的信息内容，以及根据Title、Description或Narrative的背景描述部分可能会纳入检索范围，但在Narrative域的排除条件部分明确滤除的内容。

2.2 主题功能分块

如前所述，Query形式上可分成Title、Description和Narrative等不同域。

Title和Description所用词数少，一般出现的可能是信息内容词语，不太会出现对查询操作和条件的表述词语(如“查询”“不在检索范围”)。由于字数较少，信息内容词语除了表示特定检索对象主体的专名外，还常常带有概括意义的词语(如“中国核电站情况”)，这些词语不太可能在目标文本直接出现其词形本身，往往需要对其实施一定程度的变换处理[5]。在Narrative中，除了可能存在一些表示查询本身操作的元词语外，对查询对象信息内容的描述也相应比较具体。如“电子游戏对青少年的影响”和“电子游戏对青少年学习生活行为等方面的影响，对电子游戏本身内容介绍和分析不在检索范围内”。

(1)短语描述

Title域多由简单短语构成，具有划定要检索的信息内容的作用，基本上可以当作关键词/短语检索项处理，如同Web搜索引擎接受用户在输入框中键入的检索词语[3]。Description域相比Title域的简短表述更详细，通过对Title所述主题进行检索范围或方向的限定以提升检索准确性，相应也增加了一些限制性词语。

虽然Description域比Title域描述更细致，但囿于篇幅，只有若干短语或一个简单句，并不能全面详尽地刻画用户的信息需求，因此才有Narrative域的出现。

(2)长句描述

Narrative域体现为由一组相关句子实现的一个段落。长句对信息需求的描述比较复杂，考察其内部功能，在组织结构上还有一定规则可循。如可以进一步区别以下三个表述功能块：Query背景块、核心请求块和排除条件块。图1 Narrative文本的信息内容可体现为表2形式。

表2 Narrative域中不同表述功能块的区别

语句类别	语句实例
Query背景块	垃圾邮件已经成为令每个人以及政府机构头痛的严重问题。
核心请求块	检索防范垃圾邮件的技术措施以及政府相关法律条例，
排除条件块	只介绍垃圾邮件及其危害的文章或者有关产品介绍不在检索范围内。

其中，反映查询主题范围的核心请求块不能省略，是信息需求描述的必有项；而陈述事实或交待背景的Query背景块，从反面缩小检索范围的排除条件块则为可选项。

①Query背景块

Query背景块是一种辅助语句，位于对检索目标对象详尽描述的核心请求块之前，表述用户关注信息的形成和由来，对确定检索对象范围具有一定提示作用。由于它紧靠具体查询操作语句之前，在查询条件表述语焉不详的情况下，由篇章承前连贯性质，有时需要回溯到Query背景块获取相关条件。

②核心请求块

核心请求块是具体提出检索要求的语句，紧随Query背景块后，信息内容比Title域和Description域的主题描述更详细。核心请求块经常由典型的提示表述形式显式体现，如“检索”“查询”“相关文件(应)包括/提及……”“……视为相关”等。

准确理解并提取核心请求块的主题词语是决定信息检索系统成败的关键。Narrative域是由三个功能块构成的一个段落，不同功能块间由于段落内的文本联系有连贯性，常有照应(anaphor)需要处理，因此不能只是简单地把剥离后的核心请求块中的信息内容提取出来当成关键词处理。

③排除条件块

排除条件块是明确指出需要排除的一类语句，一般在核心请求块之后，用来进一步细化检索条件。在Narrative域核

心请求块细化真正的信息需求之后，排除条件块是对检索范围的进一步限定。前者是从正例出发，后者则是从反面着眼。

如果说 Title 域或 Description 域的内容具有划定论域(Domain)或给定一个检索信息内容集 U 的作用，Narrative 域则是在此集合 U 的基础上，进一步给定限制条件：正条件(包含在核心请求块表示的检索范围的信息需求)、负条件(包含在排除条件块明确排除内容的信息需求)，从而得到更精细的信息需求表示。

正条件部分的信息内容相当于在全集取其中满足用户特定需求的某个子集；而负条件不是简单的 Not 算子(operator)，还应结合 Title 域和 Description 域框定的全集内容，即在这里，我们看到的是 But 算子[4]。

我们的服务对象是信息检索系统，而非问答系统。只要结果文本含有正条件部分要求的信息内容就可判定 Query 和文本相关，因此对目标文本中出现排除条件块的信息，可以采取较宽容的策略，即允许目标文本部分内容表述该意义，但前提是它不能成为文本的核心主题。

3 实验设计

3.1 不同主题功能块词语的加权

Title 和 Description 域的检索概念框定用户的检索对象范围，在一个相对较小的语言环境选用某些词语而不是其他词语，使得这些词语具有特殊的指别意义。根据 Query 不同域和不同功能块的词语对主题贡献度的不同，我们为其分配不同的加权方案，这是与当前向量空间系统不区分词语对主题贡献程度，统一采用词语出现频次的 TF×IDF 加权的区别之处：

$$weight\_topic(C_{title})=5$$

$$weight\_topic(C_{description})=3$$

$$weight\_topic(C_{background})=1$$

$$weight\_topic(C_{keyrequest})=2$$

$$weight\_topic(C_{exclusion})=-2$$

Title 域某一词语的每次出现，赋值为 5；Description 域每次词语出现，赋值为 3。Narrative 域中 Query 背景块(background)词语权重为 1，核心请求块(keyrequest)词语权重为 2，排除条件块(exclusion)词语权重为 −2。这是因为 Title 和 Description 域的词语大多是专指词语，在目标文本出现的可能性较大；Query 背景块由于不是检索的中心主题，故词语权值降低；而排除条件块词语多数是用户明确说明不需要的，因此加上惩罚分数。以上是根据词语在不同主题功能块对检索主题的贡献度加权。

由于同一词语可能在不同域或不同功能块都出现过，具体某一个词语的权值为其在各域中出现的权重之和。如某一词语不仅在 Title 域还在 Description 域中出现，则该词的正权值累加之后，其凸显意义得到加强；而如果一个词语仅在排除条件块出现，则为负分，其对检索主题的贡献度为负值，在文本中出现该词则有可能降低 Query 和文档的匹配值。

通过语料观察，命名实体等专指词语一般不太可能同时出现在正负条件不同的功能块中。虽然概括词语可能会跨越不同域或不同功能块，但由于概括词语本身经常不以其词形直接在文档中体现，往往由其他表述更具体意义的词语体现，因此对其判罚处理不会太大地影响 Query 和文档的相似度计算。

3.2 向量空间模型

在文档和 Query 的相似度计算实现方面，我们沿用向量空间方法作为基本实现方法。

设被检索文本记为 d_j，用户 Query 记为 q，设检索用词语数 $|q|=t$，$w_{i,j}$ 为 t 个检索词语中第 i 个词语在文档 j 中的权重，$w_{i,q}$ 为 t 个检索词语中第 i 个词语在 Query 中的权重，则有每个文档 d_j 与 Query 的相似度计算为：

$$sim(d_j,q)=\frac{\vec{d_j}\cdot\vec{q}}{|\vec{d_j}|\times|\vec{q}|}=\frac{\sum_{i=1}^{t}w_{i,j}\times w_{i,q}}{\sqrt{\sum_{i=1}^{t}w_{i,j}^2}\times\sqrt{\sum_{i=1}^{t}w_{i,q}^2}}$$

其中 t 中第 i 个词语的文档权重计算为

$$w_{i,j}=f_{i,j}\times\log\frac{N}{n_i}$$ 归一化词频因素计算为

$$f_{i,j}=\frac{freq_{i,j}}{\max_l freq_{l,j}}$$

t 中第 i 个词语的 Query 权重计算为

$$w_{i,q}=\left(0.5+\frac{0.5\times freq_{i,q}}{\max_{l,q}\times freq_{l,q}}\right)\times\log\frac{N}{n_i}$$

通过对词语在文档和 Query 中的不同权重比较，采用夹角余弦距离作为评判结果相似性的结果值。

3.3 改进的向量空间模型

根据主题凸显词语对检索结果的影响加权，把它结合到传统向量空间模型中。最根本的改变是对最初模型中 $w_{i,q}$ 计算的改进。在基础向量空间模型中，$w_{i,q}$ 的确定依然沿用的是 $TF\times IDF$ 方法，不同位置的词语没有区别。而现在对于 Query 中的每一个词语，需要再根据不同域或功能块得到不同权值。

$$w_{i,j}=weight\_topic(W_i)\times(0.5+\frac{0.5\times freq_{i,q}}{\max_{l,q}\times freq_{l,q}})\times\log\frac{N}{n_i}$$

4 实验过程及结果

我们设想，区分词语在 Query 不同域或不同功能块中出现的特性，结合传统的单纯词频计数加权方法，比不加区别把所有分词结果送入检索系统作为输入条件，应该有更好的检出效果。由于引入主题词语的不同权重机制，能够强化一些凸显概念，从而使目标文本与 Query 更好匹配。为此，我们设计了三组实验，来试图验证这一假设。

4.1 语料对象

863IR 2004 年的 Query 集和数据集，Query 集有 30 个模拟用户提问问题；本地检索数据集有 16G，全部是网页文本，文本主题多种多样。

4.2 实验方法

作为对照，我们设置三组实验。其输入分别是：

(1) 以 Query 分词结果，排除停用词后，停用词排除处理参见我们另文描述的方法[7]，直接取所有剩余词语不利用词频等权值信息作为检索项送入检索系统；

(2) 以 Query 分词结果，排除停用词后，对所有剩余词语根据其在 Query 中的词频加权后送入检索系统，计算方法参见 3.2 节；

(3) 以 Query 分词结果，排除停用词后，对所有剩余词语根据其在 Query 中不同域或不同主题功能块的属性及词频信息加权后，送入检索系统作为检索条件。其中，不同域和不同主题功能块条件属性加权的取值参见 3.1 节的介绍，相关计算方法参见 3.3 节。

4.3 实验过程

(1) 初选阶段:由于原始863IR网页数据集巨大,我们首先选择Query集的30个问题作为输入,通过带伪相关反馈的向量空间模型,从数据集获取针对每个Query的前50个返回文档,并以此作为小型测试用语料数据库。

(2) 检索阶段:采用向量空间模型进行上述三组实验。三组实验的不同之处在于参数选用,即是否采用主题词语加权和 $TF\times IDF$ 加权。最终取Query集中12个Query作为实验样本。每个Query根据与文档的夹角余弦作为相似度计算的工具,只返回前10个检索结果。

(3) 评估阶段:对每个Query返回结果,人工考察结果文档是否切合用户Query信息需求。采用TREC QA的平均倒数序和(Mean Reciprocal Rank, MRR)来评估不同方法的优劣,即对每个Query的前10个结果文档,取满足要求的文档在结果排序列表中序位的倒数和作为最终结果。该值越大,说明相关结果排在越靠前。

我们没有采用信息检索传统的查全率(recall)和查准率(precision)作为评判标准。这是因为我们手头没有标准答案集,相对16G原始语料,很难据此作出评估。同时,即使针对初选阶段构造的小型数据集,也难保证所有正确结果都已经在初选阶段被检出。

4.4 实验结果

采用以上三种加权方案对12个Qurey检索测试,得到三组不同结果,数据见表3。

表3 三种解决方案对12个Query检出结果MRR的比较

Query号	方案1	方案2	方案3	方案2较1高	方案3较1高	方案3较2高
1	1.5	1.91	1.99	27.33%	32.67%	4.19%
2	2.09	2.14	2.93	2.39%	40.19%	36.92%
3	1.12	1.05	1.92	−6.25%	71.43%	82.86%
4	0.5	0.61	0.75	22%	50%	22.95%
5	0.38	0.38	0.58	0%	52.63%	52.63%
6	0.1	1.23	1.74	1130%	1640%	41.46%
7	0.39	0.39	0.75	0%	92.31%	92.31%
8	1.2	1.2	1.13	0%	−5.83%	−5.83%
9	0.97	1.58	1.22	62.89%	25.77%	−22.78%
10	0.73	1.98	2.66	171.23%	264.38%	34.34%
11	2.25	2.35	2.29	4.44%	1.78%	−2.55%
12	1.08	1.41	2.55	30.56%	136.11%	80.85%
平均值	1.03	1.35	1.71	31.07%	66.02%	26.67%

从表3可以看出,直接使用所有非停用词(方案1)作为输入的检索效果不尽如人意,平均值为1.03;而只采用 $TF\times IDF$ 加权方式(方案2)平均值为1.35;同时利用词语在不同域和不同功能块加权及 $TF\times IDF$ 加权方式(方案3)的平均值为1.71。方案2比方案1提高31.07%,方案3比方案1提升66.02%;方案3比方案2提升26.67%。

从中发现,Query主题词语确实存在词频和主题区域的区分影响检出效果的情况。

结论 本文主要针对当前信息检索测试Query的SGML形式特征,区分不同主题描述域;对Narrative成段落的自然语言表述形式,采用正则表达式,划分Query背景块、核心请求块和排除条件块等不同主题功能块,使得Query信息内容词语能够根据描述域和功能块划分。对处于不同描述域和功能块中的词语实施不同的加权处理,结合 $TF\times IDF$ 传统词频统计方法,使得表述主题概念的Query词语得以强化。实验表明凸显Query主题词语和对信息内容词语的 $TF\times IDF$ 加权策略比单纯采用计数的 $TF\times IDF$ 加权效果更好。

参考文献

1 Baeza-Yates R, Ribeiro-Neto B, et al. Modern Information Retrieval. ACM Press, 1999

2 Dkaki T, Mothe J. Combining Positive and Negative Query Feedback in Passage Retrieval. In: Proceedings of Recherche d'Information Assistée par Ordinateur(RIAO'2004), 2004

3 Grootjen F A, van der Weide Th P. Conceptual Query Expansion:[Technical Report]. NIII-R0406. University of Nijmegen, The Netherlands, 2004

4 Navarro G. Query Languages. In: Baeza-Yates R, Ribeiro-Neto B. eds. Modern Information Retrieval. New York: ACM Press, 1999

5 Yin L. Topic Analysis and Answering Procedural Questions: [Technical Report ITRI-04-14]. University of Brighton, 2004

6 熊文新, 宋柔. 信息检索自然语言查询问句处理框架. 计算机科学, 2006(10)

7 熊文新, 宋柔. 信息检索用户查询语句的停用词过滤. 计算机工程, 2007(6)

8 张华平. 语言浅层分析与句子级新信息检测研究:[博士学位论文]. 北京: 中国科学院计算技术研究所, 2005

(上接第162页)

14 Yang H, Chua T-S, Wang S, Koh C-K. Structured use of external knowledge for event-based open domain question answering. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval. ACM Press, Toronto, Canada, 2003

15 吴平博, 陈群秀, 马亮. 基于事件框架的事件相关文档的智能检索研究. 中文信息学报, 2003(7): 25～30

16 梁晗, 陈群秀, 吴平博. 基于事件框架的信息抽取系统. 中文信息学报, 2006(20): 40～46

17 Gruber T R. Toward principles for the design of ontologies used for knowledge sharing: A translation approach to portable ontologies. Knowledge Acquisition, 1993(5): 199～220

18 Borst W N. Construction of engineering ontologies for knowledge sharing and reuse:[PhD]. University of Twente, Enschede, 1997

19 Studer R, Benjamins VR, Fensel D. Knowledge engineer, principles and methods. Data and Knowledge Engineering, 1998, 25: 161～197

20 吴强, 刘宗田, 强宇. 基于本体的知识库推理研究. 计算机应用研究, 2005, 22: 50～52, 87

21 陆汝钤. 世纪之交的知识工程与知识科学. 清华大学出版社, 2001

22 周文, 刘宗田, 陈慧琼. FCA与本体结合研究的综述. 计算机科学, 2006, 33: 8～12

23 Lin H F, Liang J M. Event-based Ontology design for retrieving digital archives on human religious self-help consulting. Institute of Electrical and Electronics Engineers Computer Society, Piscataway, NJ 08855-1331, United States, Hong Kong, China, 2005. 522～527

24 Lee C S, Chen Y-J, Jian Z-W. Ontology-based fuzzy event extraction agent for Chinese e-news summarization. Expert Systems with Applications, 2003, 25: 431～447

25 Filatova E, Hatzivassiloglou V. Event-based Extractive summarization. In: Proceedings of ACL 2004 Workshop on Summarization, 2004. 104～111

26 Daniel N, Radev D, Allison T. Sub-event based Multi-document Summarization. In: Proceeding of the HLT-NAACL 2003 Workshop on Text Summarization, 2003. 9～16

27 Li W, Xu W, Wu M, Yuan C, Lu Q. Extractive Summarization using Inter-and Intra-Event Relevance. In: Proceedings of COLING-ACL, 2006